Problem Set 4

1. Consider an MDP where actions A are vectors $(A_1, \ldots, A_n) \in \mathcal{A}^n$, for some set $\mathcal{A}$. Therefore in each time stage the number of actions to be considered is exponential in the number n of action variables. Show that this MDP can be converted into an equivalent one with $\mathcal{A}$ actions in each time stage but a larger state space. (This problem shows that complexity in the action space can be traded for complexity in the state space, which is addressed by value function approximation methods.)

2. Show that the VC dimension of the class of rectangles in $\Re^d$ is $2d$.

3. Another value function approximation algorithm based on temporal differences is called λ least squares policy evaluation (λ-LSPE). We successively approximate the cost-to-go function J^* by $J^* \approx \Phi r_k, k = 1, 2, \ldots$. Recall that $\phi(x)$ is the row vector whose ith entry corresponds to $\phi_i(x)$. Define the temporal difference relative to approximation r_k:

$$d_k(x, y) = g(x) + \alpha\phi(y)r_k - \phi(x)r_k.$$

Then λ-LSPE updates r_k based on

$$\begin{aligned}\tilde{r}_k &= \underset{r}{\operatorname{argmin}} \sum_{m=0}^{k} \left(\phi(x_m)r - \phi(x_m)r_k - \sum_{l=m}^{k} (\alpha\lambda)^{l-m} d_k(x_l, x_{l+1}) \right)^2, \\ r_{k+1} &= r_k + \gamma(\tilde{r}_k - r_k).\end{aligned}$$

The updates can be rewritten recursively as

$$r_{k+1} = r_k + \gamma B_k^{-1}(A_k r_k + b_k),$$

where

$$\begin{aligned}B_k &= \sum_{m=0}^{k} \phi(x_m)\phi(x_m)', \\ A_k &= \sum_{m=0}^{k} z_m(\alpha\phi(x_{m+1}) - \phi(x_m)), \\ b_k &= \sum_{m=0}^{k} z_m g(x_k), \\ z_m &= \sum_{l=0}^{m} (\alpha\lambda)^{m-l}\phi(x_l).\end{aligned}$$

(a) Show that

$$\begin{aligned}\lim_{k\to\infty} \mathrm{E}A_k &= A = \Phi^T D(\alpha P - I) \sum_{m=0}^{\infty} (\alpha\lambda P)^m \Phi, \\ \lim_{k\to\infty} \mathrm{E}b_k &= b = \Phi^T D \sum_{m=0}^{\infty} (\alpha\lambda P)^m g, \\ \lim_{k\to\infty} \mathrm{E}B_k &= B = \Phi^T D\Phi.\end{aligned}$$

(b) It can actually be shown that $A_k/k \to A$, $b_k/k \to b$ and $B_k/k \to B$, with probability 1, and r_k converges to $r = -A^{-1}b$. Compare r with the limiting value of r_k achieved by TD(λ).

(c) The main disadvantage of λ-LSPE is that it requires inverting matrix B_k in each iteration. Note that $B_k \in \Re^{p \times p}$, where p is the number of basis functions. However, there is an efficient incremental scheme for inverting B_k which only requires explicitly inverting scalars in each iteration.

   i. (Matrix Inversion Lemma) Show that, for all matrices M and N, $(I + MN)^{-1} = I - M(I + NM)^{-1}N$.

   ii. Propose an iterative scheme for computing B_k that only requires scalar inversion on all iterations $k = 2, 3, \ldots$ (assume that B_k is invertible for every k).